16. Bayesov klasifikator II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.4

1 Bayesov klasifikator vs. logistička regresija

- Ideja: pokazati da logistička regresija i Bayesov klasifikator izračunavaju isti $P(y|\mathbf{x})$
- Model **logističke regresije**:

$$h(\mathbf{x};\mathbf{w}) = P(y=1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

- Aposteriorna vjerojatnost za **kontinuirani Bayesov klasifikator** (za dvije klase):

$$\begin{aligned}
P(y=1|\mathbf{x}) &= \frac{p(\mathbf{x}|y=1)P(y=1)}{p(\mathbf{x}|y=1)P(y=1)+p(\mathbf{x}|y=2)P(y=2)} = \frac{1}{1+\frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}} = \\
&= \frac{1}{1+\exp\left(\ln\frac{p(\mathbf{x}|y=2)P(y=2)}{p(\mathbf{x}|y=1)P(y=1)}\right)} = \frac{1}{1+\exp(-\alpha)} = \sigma(\alpha)
\end{aligned}$$

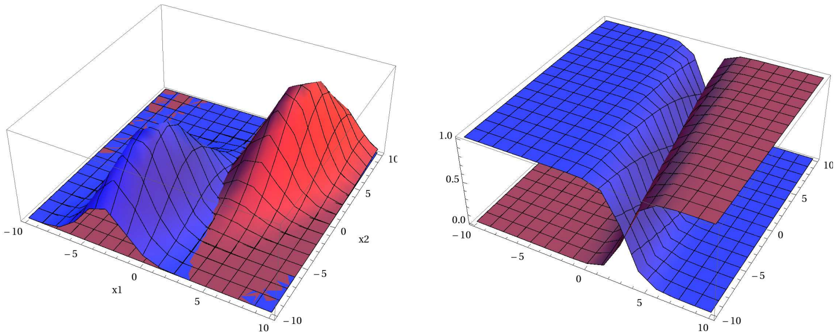
gdje

$$\alpha = \ln\frac{p(\mathbf{x}|y=1)P(y=1)}{p(\mathbf{x}|y=2)P(y=2)} = \underbrace{\ln p(\mathbf{x}|y=1)P(y=1)}_{h_1(\mathbf{x})} - \underbrace{\ln p(\mathbf{x}|y=2)P(y=2)}_{h_2(\mathbf{x})}$$

- Možemo li α prikazati kao linearnu kombinaciju težina, $\alpha = \mathbf{w}^{\mathrm{T}}\mathbf{x}$?
- Da, ako pretpostavimo **dijeljenu kovarijacijsku matricu**:

$$\begin{aligned}
\alpha &= h_1(\mathbf{x}) - h_2(\mathbf{x}) \\
&= \mathbf{x}^{\mathrm{T}}\underbrace{\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}_{\mathbf{w}}\underbrace{-\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1+\frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2+\ln\frac{P(y=1)}{P(y=2)}}_{w_0} = \mathbf{w}^{\mathrm{T}}\mathbf{x}+w_0
\end{aligned}$$

$\Rightarrow$ logistička regresija istovjetna je Bayesovom klasifikatoru s dijeljenom $\boldsymbol{\Sigma}$



- Broj parametara: $\frac{n}{2}(n+1) + 2n + 1$ (Bayes) vs. $n+1$ (logistička regresija)
  $\Rightarrow$ diskriminativan model daje istu predikciju, ali s manje parametara

2 Naivan Bayesov klasifikator

- $\mathbf{x} = (x_1, \ldots, x_n)$ ima $\prod_{k=1}^{n} K_k$ mogućih vrijednosti
- $p(\mathbf{x}|y)$ kao kategorička razdioba od $\mathbf{x} \Rightarrow$ **previše parametara** i **nema generalizacije**
- Pojednostavljenje uvođenjem **induktivnih prepostavki** u obliku uvjetnih nezavisnosti
- Pretpostavka: u svakoj klasi, svaka značajka uvjetno je nezavisna od svih drugih:

$$x_k \perp (x_1 \ldots, x_{k-1})|y \;\Leftrightarrow\; P(x_k|x_1, \ldots, x_{k-1}, y) = P(x_k|y)$$

- Faktorizacija uz tu pretpostavku:

$$P(x_1, \ldots, x_n|y) = \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_{k-1}, y) = \prod_{k=1}^{n} P(x_k|y)$$

- **Naivan Bayesov klasifikator** (*Naïve Bayes classifier*):

$$h(x_1, \ldots, x_n) = \underset{j}{\operatorname{argmax}} \; P(y=j) \prod_{k=1}^{n} P(x_k|y=j)$$

- Procjena parametara – MLE za $P(y)$ i MAP za $P(\mathbf{x}|y)$:

$$P(y=j) = \hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = j\} = \frac{N_j}{N}$$

$$P(x_k|y=j) = \hat{\mu}_{k,j} = \frac{\sum_{i=1}^{N} \mathbf{1}\big\{x_k^{(i)} = x_k \wedge y^{(i)} = j\big\} + 1}{\sum_{i=1}^{N} \mathbf{1}\{y^{(i)} = j\} + K_k} = \frac{N_{kj} + 1}{N_j + K_k}$$

  gdje je N_j broj primjera u klasi j, a N_k broj vrijednosti značajke k

- Broj parametara: $\sum_{k=1}^{n} (K_k - 1) \cdot K + K - 1$
- Pretpostavka uvjetne nezavisnosti uglavnom ne vrijedi, no model u praksi radi dobro

3 Uvjetna nezavisnost

- **Uvjetna nezavisnost** X i Y uz dani Z – notacija: $X \perp Y | Z$:

$$
\begin{aligned}
P(X, Y|Z) &= P(X|Z)P(Y|Z) \\
P(X|Y, Z) &= P(X|Z) \\
P(Y|X, Z) &= P(Y|Z)
\end{aligned}
$$

4 Polunaivan Bayesov klasifikator

- Ideja: **združiti** (ne faktorizirati) varijable koje nisu uvjetno nezavisne
- Npr., ako $x_2 \not\perp x_3 | y$:

$$P(x_1, x_2, x_3, y) = P(x_1|y)P(x_2, x_3|y)P(y)$$

  $\Rightarrow$ slabije pretpostavke $\Leftrightarrow$ složeniji model $\Leftrightarrow$ više parametara

- Broj mogućih združivanja $\Leftrightarrow$ broj particija n-članog skupa $\Leftrightarrow$ **Bellov broj** B_n
- Previše kombinacija $\Rightarrow$ **heurističko pretraživanje** na temelju **kriterija združivanja**
- Dva pristupa:
  - **točnost modela** (unakrsna provjera) – algoritam FSSJ
  - **procjena zavisnosti varijabli** – algoritmi TAN i k-DB

Algoritam FSSJ

1. Inicijaliziraj $X = \varnothing$. Početna faktorizacija:

$$P(x_1, \dots, x_n, y) = P(x_1) \cdots P(x_n)P(y)$$

$$P(y|x_1, \dots, x_n) = P(y)$$

   Klasificiraj primjere iz skupa za provjeru: $y^* = \mathrm{argmax}_j \, P(y = j)$

2. Za svaku varijablu $x_i \notin X$ koja još nije uključena u model, razmotri:
   (a) Uključi x_i kao uvjetno nezavisnu u odnosu na ostale varijable za danu klasu j
   (b) Uključi x_i tako da se ona doda u zajednički faktor s nekom već uključenom varijablom
3. Izaberi x_i i opciju koja minimizira pogrešku generalizacije
4. Ponavljaj od koraka (2) do konvergencije pogreške

- **Uzajamna informacija** – zavisnost varijabli kao odstupanje $P(x, y)$ od $P(x)P(y)$:

$$I(x, y) = D_{\mathrm{KL}}\big(P(x, y)||P(x)P(y)\big) = \sum_{x,y} P(x, y) \ln \frac{P(x, y)}{P(x)P(y)}$$

$\Rightarrow I(x, y) = 0 \Leftrightarrow x \perp y$, $I(x, y) > 0 \Leftrightarrow x \not\perp y$

- $D_{\mathrm{KL}}(P||Q)$ – **Kullback-Leiblerova divergencija** (odstupanje) distribucije P od Q:

$$D_{\mathrm{KL}}(P||Q) = \sum_{x} P(x) \ln \frac{P(x)}{Q(x)}$$

$\Rightarrow$ **relativna entropija** $P(x)$ u odnosu na $Q(x)$